Disentangling Topic Models: A Cross-cultural Analysis of Personal Values through Words

Disentangling Topic Models: A Cross-cultural Analysis of Personal Values through Words

Steven R. Wilson and **Rada Mihalcea**
University of Michigan
steverw@umich.edu, mihalcea@umich.edu

Ryan L. Boyd and **James W. Pennebaker**
University of Texas at Austin
ryanboyd@utexas.edu, pennebaker@mail.utexas.edu

Abstract

We present a methodology based on topic modeling that can be used to identify and quantify sociolinguistic differences between groups of people, and describe a regression method that can disentangle the influences of different attributes of the people in the group (e.g., culture, gender, age). As an example, we explore the concept of personal values, and present a cross-cultural analysis of value-behavior relationships spanning writers from the United States and India.

1 Introduction

Topic modeling describes a family of approaches that capture groups of related words in a corpus. In these frameworks, a *topic* can be thought of as a group of words found to be related to a higher level concept. Generally, a topic is represented as a set of numbers that describe the degree to which various words belong, which often takes the form of a probability distribution over words. Several topic modeling approaches have been proposed in the past, including Latent Dirichlet Allocation (Blei et al., 2003), Correlated Topic Models (Blei and Lafferty, 2006), Hierarchical Dirichlet Processes (Teh et al., 2012), and the Meaning Extraction Method (MEM) (Chung and Pennebaker, 2008), among others. Topic modeling has been a useful way to handle myriad tasks, including dimensionality reduction (Lacoste-Julien et al., 2009), data exploration (Blei, 2012), creation of features that are used for downstream tasks such as document classification (Zhu et al., 2009), twitter hashtag recommendation (Godin et al., 2013), and authorship attribution (Steyvers et al., 2004).

In this paper, we use topic modeling to explore sociolinguistic differences between various groups of authors by identifying groups of words that are indicative of a target process. We introduce a number of strategies that exemplify how topic modeling can be employed to make meaningful comparisons between groups of people. Moreover, we show how regression analysis may be leveraged to disentangle various factors influencing the usage of a particular topic. This facilitates the investigation of how particular traits are related to psychological processes.

We provide an example application in which we investigate how this methodology can be used to understand personal values, their relationships to behaviors, and the differences in their expression by writers from two cultures. To carry out these analyses, we examine essays from a multicultural social survey and posts written by bloggers in different countries. Our results show that culture plays an important role in the exploration of value-behavior relationships

Our contributions include: 1) a new sociolinguistic geared methodology that combines topic modeling with linear regression to explore differences between groups, while specifically accounting for the potential influence of different attributes of people in the group; 2) a cross-cultural study of values and behaviors that uses this methodology to identify differences in personal values between United States (US) and India, as well as culture-specific value-behavior links; and 3) a social survey data set containing free response text as well as a corpus of blog posts writ-

Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pages 143–152,
Austin, TX, November 5, 2016.

ten by authors from two countries.

2 Methodology

2.1 Topic Modeling with the Meaning Extraction Method

While several topic modeling methods are available, we use the MEM as it has been shown to be particularly useful for revealing dimensions of authors' thoughts while composing a document (Kramer and Chung, 2011; Lowe et al., 2013). The MEM was first used as a content analysis approach for understanding dimensions along which people think about themselves as inferred from self descriptive writing samples. Given a corpus in which the authors are known to be writing in a way that is reflective of a certain psychological construct (e.g., self concept), the MEM can be used to target that construct and automatically extract groups of words that are related to it. Note that the MEM is a general framework for identifying topics in a corpus, and is one of many approaches that could be taken toward this goal. While our methodology allows for flexibility in decision making during the process, we opt for the original MEM setting proposed in (Chung and Pennebaker, 2008) and leave the investigation of the effectiveness alternative configurations for future work.

The standard MEM begins with a particular series of preprocessing steps, which we perform using the Meaning Extraction Helper (Boyd, 2015). This tool tokenizes and lemmatizes the words in each document, then filters out function words as well as rare words (those used in less than 5% of documents). Each of the documents is then converted into a binary vector indicating the presence of a given word with a value of 1 and the absence of a word with a 0. This approach is taken in order to focus on whether or not documents contain particular words without taking into account word frequency.

Based on the notion that word co-occurrences can lead to psychologically meaningful word groupings, we then perform principal components analysis on the correlation matrix of these document vectors, and apply the varimax rotation (Kaiser, 1958),[1] which, in terms of the language analysis domain, is formulated as the orthogonal rotation that satisfies:

$$max \sum_{t}^{T} \left(\sum_{w}^{V} {f_{wt}}^4 - \frac{(\sum_{w}^{V} {f_{wt}}^2)}{|V|} \right)$$

where T represents the set of topics ($|T| = k$, the number of topics specified as a parameter to the model), V is the vocabulary of all the words in the data set, and f_{tw} is the factor loading of word (variable) w for topic (factor) t. The goal of this rotation is to increase structural simplicity and interpretability while maintaining factorial invariance.

For many topic modeling approaches, the raw membership relation m_{RAW} for a word w in a topic, or "theme", t, may be defined directly as: $m_{RAW}(t, w) = f_{wt}$ where f_{wt} is the factor loading of w for t (or posterior probability of w belonging to t, depending on the paradigm being used). However, the MEM traditionally takes a thresholding approach to words' membership to a topic: any word with a factor loading of at least .20 for a particular component is retained as part of the theme, (words with loadings of less than -.20 reflect concepts at the opposite end of a bipolar construct). Functionally, then, we define the threshold membership relation m_{THRESH} for a word w to a new theme t:

$$m_{THRESH}(t, w) = \begin{cases} 1 & \text{if } f_{wt} > \tau, \\ -1 & \text{if } f_{wt} < -\tau, \\ 0 & otherwise. \end{cases}$$

We follow (Chung and Pennebaker, 2008) and choose a threshold of $\tau = .2$.

2.2 Topic Regression Analysis

To measure the degree to which a particular topic is used more (or less) by one group than another, we fit and subsequently analyze a series of regression models. For each document d and theme t, we assign a usage score by the function:

$$s(t, d) = \frac{\sum_{w}^{d} m(t, w)}{|d|},$$

assuming that a document is an iterable sequence of words and m is the chosen membership relation. When using m_{THRESH}, this score is essentially a

[1] We use the implementation of the varimax rotation from the stats package of CRAN (cran.r-project.org).

normalized count of words in a document that belong to a particular theme minus the total number of words that were found to be in opposition to that theme (those words for which $m(t, w) = -1$).

We then regress the normalized score:

$$s_{NORM}(t, i, D) = \frac{|D| \cdot s(t, d_i)}{\sum_{d \in D} s(t, d)}$$

against variables encoding attributes of interest pertaining to each document d_i, such as the author's membership to a certain group, in order to determine the influence of these attributes on $s_{NORM}(t, i, D)$. Here, D represents all documents in the corpus and d_i is the ith document in D.

After fitting the regression models, we can interpret the coefficient attached to each attribute as the expected change in the usage of a particular theme as a result of a unit increase in the attribute, holding all other modeled attributes constant. For example, if we have a variable measuring the gender of the document's author, encoded as 0 for male and 1 for female, we can explore the degree to which gender has an expected relationship with the usage of a theme while controlling for other possible confounding factors that are included in the regression model. With this formulation, a binary variable with a predicted coefficient of, e.g., .15 would indicate an expected 15% increase in the usage of a theme between the group encoded as 1 (female, in our example) over the group encoded as 0 (male). Furthermore, we check for interactions between the attributes through a two-level factorial design regression analysis.

2.3 Relationships Between Sets of Themes

It may also be desirable to quantify the relationships between two different sets of themes. If the same set of authors have written texts that are known to relate to multiple categories of interest, perhaps psychological constructs (e.g., an essay about personality and another about mental health), the MEM can be run for each category of writing in order to generate several sets of themes.

At this point, this is equivalent to treating each writing type as a distinct meaning extraction task where the texts from a corpus C_1 generates T_1 and another corpus C_2 generates T_2, where C_1 and C_2 are collections of documents belonging to distinct categories (e.g., stances on a political issue and views of morality). We are then able to take a look at the relationships *within* or *between* the constructs as expressed in texts of C_1 and C_2. We use the previously defined s function to assign a score to each writing sample $d \in C_i$ for each topic $t \in T_i$ so that all documents are represented as vectors of topic scores, with each element corresponding to one of the k topics. Transposing the matrix made up of these vectors gives vectors for each topic with a length equal to the number of documents in the corpus. We then use these topic vectors to compute the Pearson correlation coefficient between any pair of themes. In order to ensure that correlations are not inflated by the presence of the same word in both themes, we first remove words that appear in any theme in T_1 from all themes in T_2 (or vice versa). When using an m function that gives a continuous nonzero score to (nearly) every word for every topic, it would be advisable to use a threshold in this case, rather than absence/presence. That is, remove any words from any theme $t_i \in T_1$ with $|m(t_i, w)| > \phi$ from every topic $t_j \in T_2$ for which it is also the case that $|m(t_j, w)| > \phi$, for some small value ϕ.

These quantified topical relationships are then used as a way to look at differences between two groups of people in a new way (e.g., differences between Republicans and Democrats). To illustrate, assume that we have two groups of writers, G_1 and G_2, and writers from each group have created two documents each, one belonging to C_1 and the other to C_2, on which we have applied the MEM to generate sets of themes T_1 and T_2 and computed $s(t, d)$ scores. Then, for the group G_1, we can use the aforementioned approach to compute the relationship between every theme in T_1 and every theme in T_2 and compare these relationships to those found for another group of people, G_2. Also, we are able to compute the relationships between themes that are found when combining texts from both writer groups into a single corpus (written by $G_1 \cup G_2$) and examine how these differ from the relationships found when only considering one of the groups.

Since many correlations will be computed during this process, and each is considered an individual statistical test, correction for multiple hypothesis testing is in order. This is addressed using a series of 10K Monte Carlo simulations of the gener-

ation of the resulting correlation matrix in order to compute statistical significance, following the multivariate permutation tests proposed by Yoder et al. (2004). Each iteration of this approach involves randomly shuffling the topic usage scores for every topic, then recomputing the correlations to determine how often a given correlation coefficient would be found if the usage scores of themes by a user were randomly chosen. Observed coefficient values larger than the coefficient at the $1-\alpha/2$ percentile or smaller than the coefficient at the $\alpha/2$ percentile of all simulated coefficients are labeled as significant.

3 Example Application: Personal Values

As an example application of this methodology, we take a look at the psychological construct of *values* and how they are expressed differently by people from India and people from the US. In psychological research, the term *value* is typically defined as a network of ideas that a person views to be desirable and important (Rokeach, 1973). Psychologists, historians, and other social scientists have long argued that people's basic values predict their behaviors (Ball-Rokeach et al., 1984; Rokeach, 1968); it is generally believed that the values which people hold tend to be reliable indicators of how they will actually think and act in value-relevant situations (Rohan, 2000). Further, human values are thought to generalize across broad swaths of time and culture (Schwartz, 1992) and are deeply embedded in the language that people use on a day-to-day basis (Chung and Pennebaker, 2014).

While values are commonly measured using tools such as the Schwartz Values Survey (SVS), a well established questionnaire that asks respondents to rate value items on a Likert-type scale (Schwartz, 1992), it has recently been shown that the MEM is another useful way to measure specific values, and can be applied to open-ended writing samples (Boyd et al., 2015). We show how the MEM can be used to target the concept of values to create useful themes that summarize the main topics people discuss when reflecting on their personal values in two different cultural groups. While doing this, we seek to avoid overlooking culture, which is a considerable determiner of an individual's psychology (Heine and Ruby, 2010). Importantly, research studies that focus exclusively on very specific people groups may reach false conclusions about the nature of observed effects (Henrich et al., 2010; Peng et al., 1997).

Since values are theorized to relate to a person's real-world behaviors, we also use the MEM to learn about people's recent activities and which values these activities link to most strongly within different cultural groups. Furthermore, we show how the themes that we discover can be used to study cultural value and behavior differences in a new social media data set.

4 Data Collection

4.1 Open-Ended Survey Data

We set out to collect data that captures the types of things people from the different cultural groups generally talk about when asked about their values and behaviors. To do this, we collect a corpus of writings from US and Indian participants containing responses to open-ended essay questions. The choice to use participants from both the US and India was grounded in three practical concerns. First, both countries have a high degree of participation in online crowdsourcing services. Second, English is a commonly-spoken language in both countries, making direct comparisons of unigram use relatively straight-forward for the current purposes. Lastly, considerable research has shown that these two cultures are psychologically unique in many ways (Misra and Gergen, 1993), making them an apt test case for the current approach.

We construct two sections of a social survey that is designed using Qualtrics survey software and distributed via Mechanical Turk (MTurk). Participants are asked to respond to the following prompt:

> *For the next 6 minutes (or more), write about your central and most important values that guide your life. Really stand back and explore your deepest thoughts and feelings about your basic values. [...]*

Additionally, since values are theorized to be related to real-world behaviors, we would like to collect some information about what people had been doing recently. Therefore, participants are also asked to write about their activities from the past week. The order of the two essay questions (values and behaviors) is randomized.

In order to guarantee an adequate amount of text for each user, we only retain surveys in which respondents write at least 40 words in each of the writing tasks. Additionally, each essay is manually checked for coherence, plagiarism, and relevance to the prompt. Within the survey itself, multiple "check" questions were randomly placed as a means of filtering out participants who were not paying close attention to the instructions; no surveys are used in the current analyses from participants who failed these check questions. After this filtering process, we choose the maximum number of surveys that would still allow for an equal balance of data from each country. Since there were more valid surveys from the US than from India, a random subsample is drawn from the larger set of surveys to create a sample that is equivalent in size to the smaller set. These procedures result in 551 completed surveys from each country, or 1102 surveys in total, each with both a value and behavior writing component.

In the set of surveys from India, 35% of respondents reported being female and 53% reported that they were between 26 and 34 years old. 96% reported having completed at least some college education. For the respondents from the US, 63% reported being female and 38% were between the ages of 35 and 54 (more than any other age range). 88% reported having had some college education.

4.2 Blog Data

To further explore the potential of this approach, we would like to apply our sets of themes to a naturalistic data source that is unencumbered by researcher intervention. While survey data is easily accessible and fast to collect, it may not necessarily reflect psychological processes as they occur in the real world. Thus, for another source of data, we turn to a highly-trafficked social media website, Google Blogger.[2]

We create a new corpus consisting of posts scraped from Google Blogger. First, profiles of users specifying that their country is India or the US are recorded until we have amassed 2,000 profiles each. Then, for each public blog associated with each profile (a user may author more than one blog), we collect up to 1,000 posts. Since a disproportionate number of these posts were written in more recent months, we balance the data across time by randomly selecting 1,000 posts for each country for each month between January 2010 and September 2015. This way, there should not be a bias toward a particular year or month when the bloggers may have been more active in one of the countries. Each post is stripped of all HTML tags, and the titles of the posts are included as part of the document.

[2]http://www.blogger.com

5 Results

5.1 Targeted Topic Extraction

First, we apply the MEM to the set of values essays, C_{VALUES}, from all respondents of the social survey. The set of extracted value-relevant themes, T_{VALUES}, is displayed in Table 1. The number of themes, k, is chosen for topical interpretability (e.g., in this case, $k = 15$). As with other topic modeling methods, slight variations in theme retention are possible while still reaching the same general conclusions. The theme names were manually assigned and are only for reference purposes; each theme is itself a collection of words with scores of either +1 or -1. For each theme, sample words that had a positive score are given. Note that each word may appear in more than one theme. The themes are listed in descending order by proportion of explained variance in the text data.

Table 2 shows the behavior themes (T_{BEHAV}). Most of these themes are rich in behavioral content. However, a few themes capture words used

Theme	Example Words
Respect others	people, respect, care, human, treat
Religion	god, heart, belief, religion, right
Family	family, parent, child, husband, mother
Hard Work	hard, work, better, honest, best
Time & Money	money, work, time, day, year
Problem solving	consider, decision, situation, problem
Relationships	family, friend, relationship, love
Optimism	enjoy, happy, positive, future, grow
Honesty	honest, truth, lie, trust, true
Rule following	moral, rule, principle, follow
Societal	society, person, feel, thought, quality
Personal Growth	personal, grow, best, decision, mind
Achievement	heart, achieve, complete, goal
Principles	important, guide, principle, central
Experiences	look, see, experience, choose, feel

Table 1: Themes extracted by the MEM from the values essays, along with example words.

in more of a structural role when composing a text descriptive of one's past events (for example, Days and Daily routine). The theme labeled MTurk is a byproduct of the data collection method used, as it is expected that many of those surveyed would mention spending some time on the site within the past week.

5.2 Topic Regression Analysis

As we explore the differences in theme usage between cultures, we attempt to control for the influences of other factors by adding gender (x_G) and age (x_A) variables to the regression model in addition to country (x_C):

$$y_i = \beta_0 + \beta_1 x_{Ci} + \beta_2 x_{Gi} + \beta_3 x_{Ai} + \epsilon_i$$

where $y_i = s_{NORM}(t, i, D)$ for theme t and the document in D with index i. We set the country indicative variable, x_C, equal to 0 if the author of a document is from the US, and 1 if the author is from India. $x_G = 0$ indicates male, $x_G = 1$ indicates female. x_A is binned into (roughly) 10 year intervals so that a unit increase corresponds to an age difference of about a decade with higher numbers corresponding to older ages. No significant interactions between country, gender, and age were detected at $\alpha = .05$ using level-2 interactions. The predicted regression coefficients are shown in Figure 1.

Theme	Example Words
Days	monday, tuesday, friday, sunday, today
Everyday activ.	shower, coffee, lunch, eat, sleep
Chores	clean, laundry, dish, cook, house
Morning	wake, tea, morning, office, breakfast
Consumption	tv, news, eat, read, computer
Time	week, hour, month, day, minute
Child care	daughter, son, ready, school, church
MTurk	computer, mturk, survey, money
Grooming	tooth, dress, hair, brush, shower
Video games	play, game, video, online, talk
Home leisure	television, snack, show, music, listen
Commuting	move, house, drive, work, stay
Family	sister, brother, birthday, phone, visit
Road trip	drive, meet, plan, car, trip
Daily routine	daily, regular, routine, activity, time
Completion	end, complete, finish, leave, weekend
Friends	friend, visit, movie, together, fun
Hobbies	garden, read, exercise, write, cooking
School	attend, class, work, project, friend
Going out	shop, restaurant, food, family, member
Taking a break	break, fast, chat, work, routine

Table 2: Themes extracted by the MEM from the behavior essays, along with example words.

Even when using the same set of topics, we see cultural differences coming into play. Culture coefficients for the value themes show that Hard work and Respect for others were predominately talked about by Americans. Indian authors tended to invoke greater rates of the Problem Solving, Rule Following, Principles, and Optimism themes. The theme containing words relating to the value of one's Family had a significant coefficient indicating that it is generally used by females more than males.

5.3 Value-behavior Relationships

Next, we look at how usage of words from the value themes relates to usage of words from the behavior themes. Table 3 shows the correlations between topics in T_{VALUES} and T_{BEHAV}. These correlations were computed three times: once each for texts written by only people from India, texts written by only by people from the US, and for the entire set of texts. Overall, all but three of the behavior themes have observable links to the values measured in at least one of the cultural groups.

Looking more closely at the results, we see that only one of the value-behavior relationships is shared by these two cultures: the value of Family is positively related to the behavior Child care. This result is also identified when looking at the combination of texts from both cultures. One potential explanation for this is that, as we have shown, the use of words from the Family theme is more related to a person's gender than her/his culture, so removing texts from one culture will not affect the presence of this relationship. On the other hand, when considering only the text from American survey respondents, we notice that the value of Hard work is related to Chores. However, if we ignored these writing samples and only analyzed the texts from Indian authors, we saw that this same theme of Hard work is related to Consumption and Home leisure. The combined set of texts captures all three relationships. This may hint at the solution of simply combining the texts in the first place, but further investigation showed that some of the relationships only emerged when examining texts from a single country. For example, we would not learn that American authors who

	Respect others	Religion	Family	Hard work	Time & money	Problem solving	Relationships	Optimism	Honesty	Rule following	Societal	Personal growth	Achievement	Principles	Experiences
Days															
Everyday activities													●		
Chores				●♦	◊	◊									
Morning				◊		♦		♦		♦				♦	
Consumption				■♦									●		
Time	○														
Child Care			●■♦				●♦								
MTurk				♦				◊							
Grooming													●		
Video games				♦											
Home leisure				■♦											
Commuting						◊		□◊						◊	
Family	●													◊	
Road trip				●											
Daily routine	◊			◊		♦		●♦		♦					
Completion															
Friends											●				
Hobbies	●												●		
School		◊				♦		♦						♦	
Going out									■						
Taking a break															

Table 3: Coverage of behavior MEM themes (rows) by value MEM themes (columns) for two different cultures. All results significant at $\alpha = .05$ (two-tailed).

USA only: ● : $r > 0$, ○ : $r < 0$, **India only**: ■ : $r > 0$, □ : $r < 0$, **Combined**: ♦ : $r > 0$, ◊ : $r < 0$

wrote about Achievement in their values essay were more likely to have talked about Personal Grooming when listing their recent activities, or that Indian authors who used words from the value theme of Honesty probably wrote more words from the Going Out theme.

5.4 Applying Themes to Social Media Data

For the blog data, C_{BLOGS}, we perform topic modeling procedures that are parallel to those described earlier, with one exception: due to an extreme diversity in the content of blog posts, the threshold at which rare words were removed was set to 1% in order to capture a greater breadth of information. We found that a large number of themes (nearly 60) was required in order to maximize interpretability and keep unrelated topics from mixing. Spatial limitations preclude the presentation of all themes in the current paper, therefore, we present those themes that were later found to be most related to personal values in Table 4.[3]

Since value-relevant themes, T_{VALUES}, were established using the MEM on the value survey essays, value-specific language can be captured in the blog data without the need for a separate MEM procedure to be conducted. Themes in Table 4, then, reflect a broader, more naturalistic set of concepts being discussed by bloggers in the real world (T_{BLOGS}) that can then be linked with their value-relevant language as measured by computing $s(d, t)$ for $d \in C_{BLOGS}$ and $t \in S_{VALUES}$. As was done in the value-behavior comparison using only the survey data, all words that appeared in any value theme were removed from all of the blog themes so that relationships were not confounded by predictor/criterion theme pairs containing overlapping sets of words. We present the themes found when looking at blog posts from each culture individually as well as the

[3]A complete list of themes and unigram loadings are available from the first author by request.

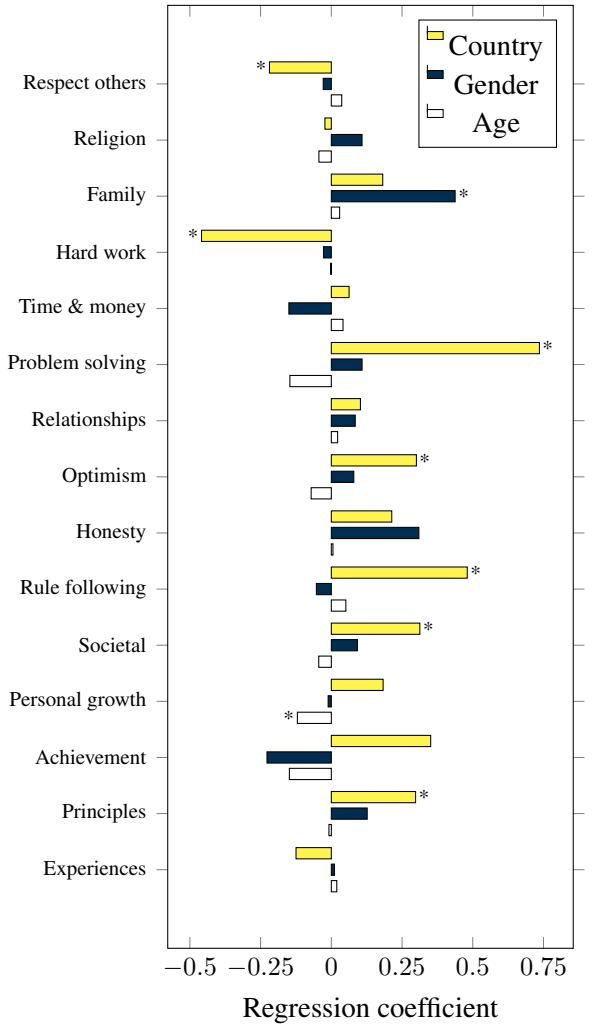

Figure 1: Coefficients for the Country, Gender, and Age variables in regression model. For Country, Gender, and Age, negative values indicate a US, male, or younger bias toward the theme, respectively, and positive values indicate an Indian, female, or older bias toward the theme, respectively. * indicates $p < .001$.

full combined corpus in Table 5.

In this dataset, we saw a similar trend as in Table 3: the particular cultural composition of the corpus changes the observed relationships. However, the association between the Religion 1 blog theme and the Religion, Honesty, and Experiences value themes was present in both US and India when considered in isolation, as well as in the combined corpus. The Tech industry theme was negatively correlated with a large number of value themes, which alludes to the idea that the words in this theme are actually an indicator of less value-related language in general. Many of the relationships found in one of the cultures were also found using the combined corpus, but only in the US data did we see a significant increase in respectful language for blogs talking about the environment; only in India did we find a negative relationship between the value theme of Personal growth and posts about the Stock market.

6 Conclusions

We have presented a methodology that can be used to employ topic models to the understanding of sociolinguistic differences between groups of people, and to disentangle the effects of various attributes on a person's usage of a given topic. We showed how this approach can be carried out using the MEM topic modeling method, but leave the framework general and open to the use of other topic modeling approaches.

As an example application, we have shown how topic models can be used to explore cultural differences in personal values both qualitatively and quantitatively. We utilized a open-ended survey as well

Theme	**Example Words**
Religion 1	jesus, glory, saint, angel, pray
Outdoorsman	farm, hunt, wild, duty, branch
Government	government, department, organization
Religion 2	singh, religion, praise, habit, wise
Profiles	french, russian, male, female, australia
Personal life	cry, job, sleep, emotion, smile
Financial	sector, money, trade, profit, consumer
School	school, university, grade, teacher
Stock market	trade, market, close, investor, fund
Tech industry	software, google, microsoft, ceo
Sports	league, play, win, team, score
Cooking	recipe, delicious, prepare, mix, kitchen
US Politics	washington, obama, debt, law, america
Job openings	requirement, candidate, opening, talent
Crime	murder, police, crime, incident
Film industry	direct, film, movie, actor, musical
India & China	india, china, representative, minister
Space exploration	mars, mission, space, flight, scientist
Environment	weather, earth, bird, storm, ocean
Indian city living	delhi, financial, tax, capital, chennai
Beauty	gold, pattern, hair, mirror, flower
Happy fashion	clothes, funny, awesome, grand

Table 4: Sample themes extracted by the MEM from the blog data, along with example words.

	Respect others	Religion	Family	Hard work	Time & money	Problem solving	Relationships	Optimism	Honesty	Rule following	Societal	Personal growth	Achievement	Principles	Experiences
Religion 1		●■♦	♦	◊					●■♦				♦	♦	○□◊
Outdoorsman				●♦							●♦	●♦			
Government	◊				◊		□◊	□◊			□◊				
Religion 2									■♦						
Profiles		□◊					■♦								
Personal life				●♦	■♦		♦			◊	●♦	●♦			
Financial	□◊				◊		○□◊		□◊		◊			□◊	
the School			■♦												
Stock market			◊				○	■♦	□◊		◊	□			
Tech industry	○◊	◊	○□◊		○◊	□◊	○□◊	○□◊	□◊		○□◊			○□◊	○
Sports		◊		■♦	■			◊			○◊				
Cooking	○				○										
US politics				◊			◊				□◊				◊
Job openings		□◊						□							■♦
Crime					○◊		○◊								
Film industry	□◊				○	□◊			◊	□				○◊	○
India + China							◊	□◊			◊				
Space exploration		□◊	□◊		◊				□		◊				
Indian city living	◊	□◊	□			□					◊			□	◊
Environment	●														
Beauty								◊							
Happy fashion								●	■	○◊					

Table 5: Coverage of blog MEM themes (rows) by value MEM themes (columns) for two different cultures. Correlations significant at $\alpha = .05$ (two-tailed) are presented.

USA only: ● : $r > 0$, ○ : $r < 0$, **India only**: ■ : $r > 0$, □ : $r < 0$, **Combined**: ♦ : $r > 0$, ◊ : $r < 0$

as a new collection of blog data.[4] The topics extracted from these texts by the MEM provide a high level descriptive summary of thousands of writing samples, and examining regression models gives insight into how some topics are used differently in US and India. We found that the underlying culture of the group of writers of the text has a significant effect on the conclusions that are drawn, particularly when looking at value-behavior links. In the future, we hope to explore how well culture-specific themes are able to summarize texts from the cultures from which they are derived in comparison with themes that were generated using texts from many cultures. While we focused on differences between Indian and American people, the proposed approach could also be used to understand differences in topic usage between members of any groups, such as liberals vs. conservatives, computer scientists vs. psychologists, or at-risk individuals vs. the general population.

[4]The survey data as well as the code used to download the blogs along with the list of profile URLs are available from the first author upon request.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (#1344257), the John Templeton Foundation (#48503), the Michigan Institute for Data Science, and the National Institute of Health (#5R01GM112697-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, the Michigan Institute for Data Science, or the National Institute of Health. Finally, we would like to thank Konstantinos Pappas for providing the code used to collect the blog data.

References

Sandra Ball-Rokeach, Milton Rokeach, and Joel W. Grube. 1984. *The Great American Values Test: Influencing Behavior and Belief Through Television*. Free Press, New York, New York, USA.

David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Ryan L Boyd, Steven R Wilson, James W Pennebaker, Michal Kosinski, David J Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In *Ninth International AAAI Conference on Web and Social Media*.

Ryan L. Boyd. 2015. MEH: Meaning Extraction Helper (Version 1.4.05) [Software] Available from http://meh.ryanb.cc.

Cindy K. Chung and James W. Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, 42:96–132.

Cindy K Chung and James W Pennebaker. 2014. Finding values in words: Using natural language to detect regional variations in personal concerns. In *Geographical psychology: Exploring the interaction of environment and behavior*, pages 195–216.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 593–596. International World Wide Web Conferences Steering Committee.

Steven J Heine and Matthew B Ruby. 2010. Cultural psychology. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):254–266.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Henry F Kaiser. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Adam DI Kramer and Cindy K Chung. 2011. Dimensions of self-expression in facebook status updates. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. 2009. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904.

Robert D Lowe, Derek Heim, Cindy K Chung, John C Duffy, John B Davies, and James W Pennebaker. 2013. In verbis, vinum? relating themes in an open-ended writing task to alcohol behaviors. *Appetite*, 68:8–13.

Girishwar Misra and Kenneth J. Gergen. 1993. On the place of culture in psychological science. *International Journal of Psychology*, 28(2):225.

Kaiping Peng, Richard E Nisbett, and Nancy YC Wong. 1997. Validity problems comparing values across cultures and possible solutions. *Psychological methods*, 2(4):329.

Meg J. Rohan. 2000. A Rose by Any Name? The Values Construct. *Personality and Social Psychology Review*, 4(3):255–277.

Milton Rokeach. 1968. *Beliefs, Attitudes, and Values.*, volume 34. Jossey-Bass, San Francisco.

Milton Rokeach. 1973. *The nature of human values*, volume 438. Free press New York.

Shalom H. Schwartz. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. *Advances in Experimental Social Psychology*, 25:1–65.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2012. Hierarchical dirichlet processes. *Journal of the american statistical association*.

Paul J Yoder, Jennifer Urbano Blackford, Niels G Waller, and Geunyoung Kim. 2004. Enhancing power while controlling family-wise error: an illustration of the issues using electrocortical studies. *Journal of Clinical and Experimental Neuropsychology*, 26(3):320–331.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.

Jun Zhu, Amr Ahmed, and Eric P Xing. 2009. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th annual international conference on machine learning*, pages 1257–1264. ACM.